A LASSO-Penalized BIC for Mixture Model Selection

Sakyajit Bhattacharya and Paul D. McNicholas*

Department of Mathematics & Statistics, University of Guelph.

Abstract

The efficacy of family-based approaches to mixture model-based clustering and classification depends on the selection of parsimonious models. Current wisdom suggests the Bayesian information criterion (BIC) for mixture model selection. However, the BIC has well-known limitations, including a tendency to overestimate the number of components as well as a proclivity for, often drastically, underestimating the number of components in higher dimensions. While the former problem might be soluble through merging components, the latter is impossible to mitigate in clustering and classification applications. In this paper, a LASSO-penalized BIC (LPBIC) is introduced to overcome this problem. This approach is illustrated based on applications of extensions of mixtures of factor analyzers, where the LPBIC is used to select both the number of components and the number of latent factors. The LPBIC is shown to match or outperform the BIC in several situations.

1 Introduction

Consider n realizations $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ of a p-dimensional random variable $\mathbf{X}$ that follows a G-component finite Gaussian mixture model. The likelihood is given by

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{1}$$

where $\pi_g > 0$, with $\sum_{g=1}^{G} \pi_g = 1$, are mixing proportions, $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is multivariate Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, and $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G)$. A model-based clustering approach assumes that each component or some combination of components corresponds to a cluster. When fitting the model in (1), the main task is to decide the number of components G. Titterington et al. (1985), McLachan and Basford (1988) and McLachan and Peel (2002) extensively reviewed mixture models, with a focus on Gaussian mixture models. Fraley and Raftery (2002) presented a review of work on Gaussian mixtures with a focus on clustering, discriminant analysis, and density estimation. They discuss a family of Gaussian mixture models, which arises from the imposition of constraints upon an eigen-decomposition of the component covariance structure. The family of mixture models they discuss, known as MCLUST, is actually a subset of the Gaussian parsimonious clustering models (GPCMs) of Celeux and Govaert (1995). When using the MCLUST models, one must choose the appropriate member of the family, i.e., the covariance structure, in addition to deciding the number of components G.

Ghahramani and Hinton (1997) introduced a mixture of factor analyzers model, which was further developed by Tipping and Bishop (1999) and McLachlan and Peel (2000). Through foisting constraints on the covariance structure, McNicholas and Murphy (2008, 2010) develop mixtures of factor analyzers into a

*Department of Mathematics & Statistics, University of Guelph, Guelph, Ontario, N1G 2W1, Canada. E-mail: paul.mcnicholas@uoguelph.ca.

family of parsimonious Gaussian mixture models (PGMMs). Now, in addition to selecting the member of the family (i.e., the covariance structure) and the number of components, one must also select the number of latent factors. Further complicating the model selection problem here is the fact that PGMMs are often applied to high-dimensional data. McNicholas et al. (2010) explain why the PGMMs are particularly suited to the analysis of high-dimensional data: amongst the most salient points is the fact that, unlike families like MCLUST, the number of covariance parameters is linear in data dimensionality for every member of the PGMM family.

There are a number of well-known methods to select the best mixture model but the BIC remains by far-and-away the most popular. We have

$$\text{BIC} = 2\log\mathcal{L}(\hat{\boldsymbol{\vartheta}} \mid \mathbf{x}) - \rho\log n, \tag{2}$$

where $\hat{\boldsymbol{\vartheta}}$ is the MLE of $\boldsymbol{\vartheta}$, $\mathcal{L}$ is the likelihood, ρ is the number of free parameters and n is the number of observations. For a family of mixture models, the model having the maximum BIC is selected. The use of BIC is theoretically justified by a number of authors, e.g., Kass and Wasserman (1995), Kass and Raftery (1995), and Keirbin (2000). In particular, the BIC has some useful asymptotic properties, e.g., the criterion consistently chooses the right model under an increasing number of observations (Shibata, 1986).

Nevertheless, the BIC is not without drawbacks. The criterion is derived using a Laplace approximation and its precision is influenced by the specific form of the prior density of the parameters as well as the correlation structure between observations. Recently, Clyde et al. (2007) have rectified the problems of the marginal distribution of the parameter, caused by the Laplace approximation. In addition, Fraley and Raftery (2007) proposed a Bayesian regularization for Gaussian mixtures. Their method assumes predefined priors that lead to a modified version of the BIC, using posterior modes instead of the maximum likelihood estimates (MLEs) of the parameters. The resulting method avoids degeneracies, singularities, and the problem of flat priors. However, another more serious problem has not been addressed, i.e., the problem of high-dimensional cases.

The penalty term in the BIC is $\rho\log n$, cf. (2). Therefore, in a high-dimensional setting, where $p \gg n$, the penalty term dominates the likelihood and so the BIC is prone to under fitting. Parametric estimation for high-dimensional cases has been studied by a number of authors, mostly within the linear regression set-up. The celebrated LASSO method (Tibshirani, 1996) is perhaps the most popular among them. This method minimizes the residual sum of squares under the constraint that the sum of the absolute values of the regression coefficients is less than some constant, leading to sparse solutions of the coefficients and thus an interpretable model. In the following years, different variations of the LASSO have been proposed depending on the nature of regression and asymptotic behaviour. Some of them are the adaptive LASSO (Zou, 2006), the fused LASSO (Tibshirani et al., 2005), and the graphical LASSO (Friedman et al., 2008). Fan and Li (2001) provided a theoretical discussion of variable selection via a non-concave penalized likelihood procedure where the LASSO is a special case. They also proposed that a good penalized estimation should satisfy the oracle properties, i.e., it should be consistent and the estimates should be asymptotically Gaussian.

Following the idea of Fan and Li (2001), Khalili and Chen (2007) were the first to propose the use of the penalized likelihood in finite mixture of regression models, where the penalty is non-concave LASSO being a special case. They also devised a method of selecting the tuning parameter as well as conditions under which the estimation procedure would satisfy the oracle properties. Their method is especially suitable for finite mixtures of regression models, though no new model selection criterion was proposed. It should also be noted that the theoretical results regarding the asymptotic properties were somehow strange, because the authors used the same tuning parameter comparing two different estimates for a fixed cluster. Chen and Chen (2008) proposed an extended BIC for regression in high-dimensional setting. The extended BIC assumes a prior inversely proportional to the size of the assumed model instead of a flat prior. The criterion is consistent and computationally cheap. Interestingly, the authors did not propose any penalized likelihood here, instead they maximized the natural likelihood, thus using the conventional estimation procedure. The above estimation procedures, though interesting and useful, are mainly for regression-type problems, and not applicable to mixture model-based clustering and classification. Also, as the authors rightly pointed out, the approach is computationally infeasible if $p \gg n$. Nevertheless, useful extensions can be possible. Herein, we draw upon some mathematical results from Fan and Li (2001) and Khalili and Chen (2007), especially on the issues of the choice of penalty and consistency.

The use of penalized likelihood in mixture model-based clustering has been proposed by Pan and Shen (2007), where a LASSO-type penalty is applied to the likelihood. From there, they went on to propose a modified BIC which would be well-suited for high-dimensional settings. The limitation of that method is that this criterion works only for a common, diagonal component covariance matrix. Furthermore, the authors did not study the asymptotic properties, which are important in the sense that the classical LASSO method can be inconsistent (cf. Zou, 2006). An ideal criterion should be analytically derivable from the penalized likelihood, work well for an arbitrary model, and have some good asymptotic properties. The work presented herein attempts to address these requirements by proposing LASSO-penalized BIC (LPBIC) for model selection within high-dimensional setting for the PGMM family.

While deriving the MLE of the unknown parameters, we use a penalized likelihood approach. In particular, instead of maximizing the likelihood $\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x})$, we maximize the penalized log-likelihood

$$\log \mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}) - \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p} \varphi(\mu_{gj}).$$

We use a LASSO-like penalty for $\varphi(\mu_{gj})$. In particular, $\varphi(\mu_{gj}) = n\lambda_n|\mu_{gj}|$, where μ_{gj} is the jth element in $\boldsymbol{\mu}_g$ and λ_n is the tuning parameter that depends on n. Though a LASSO penalty is used here, other types of non-concave penalties can also be suitable. For example, one might use the HARD penalty $\varphi(\mu_{gj}) = [\lambda_n^2 - (\sqrt{n}\mu_{gj} - \lambda_n)^2 \, I\left(\sqrt{n}\mu_{gj} < \lambda_n\right)]$ or the SCAD penalty, as discussed by Fan and Li (2001). One problem with using such an L_1-norm penalty is that the oracle properties might not be satisfied fully: the estimation can be consistent but not asymptotically normal. HARD or SCAD penalties satisfy both these properties and these issues are discussed in more detail in Section 3. Still, however, we prefer the LASSO-type penalty because it is computationally easier due to its convexity. From this penalized likelihood, we derive a model selection criterion. We use a modified AECM algorithm (McLachan and Peel, 2002) to estimate the parameters in the PGMM models. We show that in high-dimensional settings, our LPBIC generally outperforms the BIC for the PGMM family.

The remainder of this paper is laid out as follows. In Section 2, we discuss parameter estimation under the penalized likelihood approach and derive an LPBIC. The asymptotic properties of LPBIC are discussed (Section 3) and we illustrate our approach on real and simulated data (Section 4). The real data considered exhibit the 'small n, large p' property and our data analysis results are compared with the BIC. The paper concludes with a discussion (Section 5), while the mathematical derivation of LPBIC as well as its asymptotic properties are discussed in appendices.

2 Method

Again, suppose we observe $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ with $f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, where $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is multivariate Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. Now, instead of maximizing the likelihood $\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x})$, we maximize the penalized log-likelihood

$$\log \mathcal{L}_{\text{pen}}(\boldsymbol{\vartheta} \mid \mathbf{x}) = \log \mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}) - n\lambda_n \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p} |\mu_{gj}|, \tag{3}$$

where $\boldsymbol{\mu}_k$ and λ_n are defined as before. Hereafter, we denote $\varphi(\boldsymbol{\mu}) = \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p} \varphi(\mu_{gj})$ and so

$$\log \mathcal{L}_{\text{pen}}(\boldsymbol{\vartheta} \mid \mathbf{x}) = \log \mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}) - n\lambda_n \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p} \varphi(\mu_{gj}) = \log \mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}) - n\lambda_n \varphi(\boldsymbol{\mu}).$$

Before going into details of parameter estimation, we make two assumptions. Firstly, as we can observe, the penalty function is non-concave and singular at the origin; it does not have second derivative at 0. We locally approximate the penalty by a quadratic function as suggested by Fan and Li (2001). The parameters are estimated by successive iterations. Suppose $\boldsymbol{\mu}^{(m)}$ is the estimate of of $\boldsymbol{\mu}$ after m iterations. The penalty can be locally approximated as

$$\varphi(\boldsymbol{\mu}) \approx n\lambda_n \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p_g} \mid \mu_{gj}^{(m)} \mid + \frac{1}{2} \frac{\text{sign}\{\mu_{gj}^{(m)}\}}{\mu_{gj}^{(m)}} (\mu_{gj}^2 - \mu^{(m)}{}^2_{gj}), \tag{4}$$

where p_g is the number of non-zero elements in $\boldsymbol{\mu}_g$. We assume that the marginal distribution of the mixing proportions $(\pi_1, \pi_2, ..., \pi_g)$ is uniform on the simplex and that $\boldsymbol{\mu}_g \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_g, I(\hat{\boldsymbol{\mu}}_g)^{-1})$, for $g = 1, 2, ..., G$, where $\hat{\boldsymbol{\mu}}_g$ is the MLE derived by maximizing the penalized likelihood $\mathcal{L}_{\text{pen}}$ and $I(\hat{\boldsymbol{\mu}}_g)$ is the unit information matrix at $\hat{\boldsymbol{\mu}}_g$.

To estimate the parameters, we use the Alternating Expectation Conditional Maximization (AECM) algorithm. There are two stages of the algorithm. At the first stage of the algorithm, when estimating π_g and $\boldsymbol{\mu}_g$, we define $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$ to be indicator variables showing the component membership of the ith observation so that $z_{ig} = 1$ if $\mathbf{x}_i$ belongs to the gth component and $z_{ig} = 0$ otherwise. $\mathbf{z}_i$ is treated as the missing data at the first stage. Hence the expected complete data log-likelihood is

$$Q(\pi, \boldsymbol{\mu}) = \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \log \pi_g + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \log \left\{ \phi \left(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \right) \right\} - \varphi(\boldsymbol{\mu}),$$

where $\hat{z}_{ig} = \hat{\pi}_g \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g) / \sum_{j=1}^{G} \hat{\pi}_j \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)$. The M-step maximizes Q to update the parameter estimates π_g and $\boldsymbol{\mu}_g$. The estimation of π_g is complicated and has a complex analytic form. However, we have observed that in practical applications, the analytical estimate is equivalent to the estimate derived by the EM algorithm. Hence, in our analyses (Section 4), π_g can be estimated via

$$\hat{\pi}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig}}{n}.$$

For the mean parameters,

$$\frac{\partial Q}{\partial \boldsymbol{\mu}_g} = \hat{\boldsymbol{\Sigma}}_g^{-1} \sum_{i=1}^{n} \hat{z}_{ig} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g) - n\lambda_n \hat{\pi}_g \text{sign}(\hat{\boldsymbol{\mu}}_g).$$

Hence

$$\hat{\mu}_{gj} = \begin{cases} \text{sign}(\tilde{\mu}_{gj}) \left[|\tilde{\mu}_{gj}| - \lambda_n \left(\hat{\boldsymbol{\Sigma}}_g \mathbf{1} \right)_j \right]_+ & \text{if } \left(\hat{\boldsymbol{\Sigma}}_g \mathbf{1} \right)_j > 0, \\ \tilde{\mu}_{gj} & \text{otherwise.} \end{cases}$$

where $\tilde{\mu}_{gj} = \sum_{i=1}^{n} \hat{z}_{ig} x_{ig} / \sum_{i=1}^{n} \hat{z}_{ig}$ is the update of μ_{gj} if no penalty term were involved, $\mathbf{1}$ is the vector with every element equal to 1, and for any α, $\alpha_+ = \alpha$ if $\alpha > 0$ and $\alpha_+ = 0$ otherwise. $\hat{\mu}_{gj}$ is a shrunken estimate of μ_{gj} in the sense that $\hat{\mu}_{gj} = 0$ if $(\hat{\boldsymbol{\Sigma}}_g \mathbf{1})_j \geq 0$ and $\lambda_n > \tilde{\mu}_{gj} / (\hat{\boldsymbol{\Sigma}}_g \mathbf{1})_j$. Otherwise, $\hat{\mu}_{gj}$ is obtained by shrinking the usual EM estimate $\tilde{\mu}_{gj}$ by the amount $\lambda_n (\hat{\boldsymbol{\Sigma}}_g \mathbf{1})_j$ towards 0.

At the second stage of the AECM algorithm, we take the missing data as the group labels $\mathbf{z_i}$ and the unobserved latent factors $\mathbf{u}$ to estimate the variance-covariance matrix under the PGMM set-up. The component covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G$ are updated as usual, depending on the family of models used; see McNicholas and Murphy (2008, 2010) for details in the case of the PGMM family. The first stage, where the $\boldsymbol{\mu}_g$ and π_g are estimated based on the complete data $(\mathbf{x}, \mathbf{z})$, and the second stage, where the constituent parts of the $\boldsymbol{\Sigma}_g$ are estimated based on the complete data $(\mathbf{x}, \mathbf{z}, \mathbf{u})$, are iterated until convergence. Extensive details on an AECM algorithm for fitting the members of the PGMM family are given by McLachlan and Peel (2000) and McNicholas et al. (2010).

To derive a model selection criterion from the penalized log-likelihood, we maximize (3). Using (4), the second term of (3) becomes

$$\frac{\lambda_n}{G}\sum_{g=1}^{G}\sum_{j=1}^{p_g}\left[|\hat{\mu}_{gj}| + \frac{1}{2}\frac{\text{sign}(\hat{\mu}_{gj})}{\hat{\mu}_{gj}}(\mu_{gj}^2 - \hat{\mu}_{gj}^2)\right],$$

where p_g is the number of non-zero mean components in class g. Here we make an assumption that for a given model, the mixture components are chosen independently so that the parameters for any two clusters are independent. Hence, using the Weak Law of large Numbers with the BIC-type approximation to $\log \mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x})$, the penalized BIC is

$$\text{LPBIC} = 2\log\mathcal{L}(\hat{\boldsymbol{\vartheta}} \mid \mathbf{x}) - \tilde{\rho}\log n - \frac{2n\lambda_n}{G}\sum_{g=1}^{G}\sum_{j=1}^{p_g}\left[|\hat{\mu}_{gj}| + \frac{\left(I(\hat{\boldsymbol{\mu}}_g)^{-1}\right)_{jj}}{|\hat{\mu}_{gj}|} - \text{sign}\left(\hat{\mu}_{gj}\right)\right], \tag{5}$$

where $\tilde{\rho}$ is the number of estimated parameters which are non-zero. Intuitively, the LPBIC further penalizes the traditional BIC by both aboslute mean and absolute coefficient of variation of the parameters. The derivation is discussed in detail in Appendix A.

3 Asymptotic Properties

3.1 Properties

The consistency of a model selection criterion is closely related to the asymptotic identifiability of the model. In general, a model $\mathcal{G}$ with the the parameter set $\boldsymbol{\vartheta}$ is called identifiable if, for any two different sets of parameters $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$,

$$\mathcal{G}\left(\boldsymbol{\vartheta}_1\right) = \mathcal{G}\left(\boldsymbol{\vartheta}_2\right) \quad \Longrightarrow \quad \boldsymbol{\vartheta}_1 = \boldsymbol{\vartheta}_2.$$

We assume that our model satisfies the asymptotic identifiability condition. In the context of mixture models, a criterion is consistent if it can correctly select the number of components and the true set of parameters. If the true parameter set $\boldsymbol{\vartheta}_0$ is decomposed as $(\boldsymbol{\vartheta}_{01}, \boldsymbol{\vartheta}_{02})$ such that $\boldsymbol{\vartheta}_{02}$ contains only the zero elements, and if any estimated parameter $\hat{\boldsymbol{\vartheta}}$ that is sufficiently close to $\boldsymbol{\vartheta}_0$ is likewise decomposed as $(\hat{\boldsymbol{\vartheta}}_1, \hat{\boldsymbol{\vartheta}}_2)$, then in order to satisfy consistency, we should have $\text{P}(\hat{\boldsymbol{\vartheta}}_2 = \mathbf{0}) \longrightarrow 1$ as $n \longrightarrow \infty$ and $\hat{\boldsymbol{\vartheta}}_1 \longrightarrow \boldsymbol{\vartheta}_{01}$ in probability. Thus, the criterion should choose as it would if the true number of clusters and the true parameters were known. Based on this idea, we study the consistency of LPBIC with the help of the following assumptions:

I Let $p = \mathcal{O}\left(n^{\alpha}\right)$ and $\lambda_n = o\left(\log n/n\right)$. Define an estimate $\hat{\boldsymbol{\vartheta}}$ of $\boldsymbol{\vartheta}$ be such that $\|\ \hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\ \| = \mathcal{O}\left(n^{\kappa}\right)$ for $\kappa > -\infty$.

II Let $\boldsymbol{\vartheta} = (\theta_1, \theta_2, ..., \theta_{\nu})$. Then there exist finite real numbers M_1 and M_2 (possibly depending on κ) such that

$$\sup_j\left|\frac{\partial\log\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x})}{\partial\theta_j}\right| \leq M_1(\mathbf{x}) \quad \text{and} \quad \sup_{j,k}\left|\frac{\partial^2\log\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x})}{\partial\theta_j\partial\theta_k}\right| \leq M_2(\mathbf{x}).$$

III $I(\boldsymbol{\vartheta})$ is positive-definite for all $\boldsymbol{\vartheta}$.

Then, under assumptions I to III, and assuming that the asymptotic identifiability condition is satisfied, we state the following theorem. The proof is given in Appendix B.

Theorem If $\kappa < \min\left[0, (\alpha - 1)/2\right]$, then the LPBIC chooses the number of components and set of parameters as it would choose if $\boldsymbol{\vartheta}_0$ were known as $n \longrightarrow \infty$. In other words, under the condition $\kappa < \min\left(0, \alpha - 1/2\right)$, if there exists an estimate $\tilde{\boldsymbol{\vartheta}}$ such that $||\tilde{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0|| = \mathcal{O}\left(n^{\kappa}\right)$ and LPBIC$(\tilde{\boldsymbol{\vartheta}}) \geq$ LPBIC$(\boldsymbol{\vartheta})$ for all $\boldsymbol{\vartheta}$ such that $||\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0|| = \mathcal{O}\left(n^{\kappa}\right)$, then

Part a $\mathrm{P}(\tilde{\boldsymbol{\vartheta}}_2 = \mathbf{0}) \longrightarrow 1$ as $n \longrightarrow \infty$, and

Part b $\tilde{\boldsymbol{\vartheta}}_1 \longrightarrow \boldsymbol{\vartheta}_{01}$ in probability as $n \longrightarrow \infty$.

We prove only Part a with some of the arguments proposed by Khalili and Chen (2007) for an FMR setting. The method is modified for mixture models with high-dimensional set-up. Part b of the theorem can be proved exactly by the method described in Fan and Li (2001). To prove Part b, we need $\sqrt{n}\lambda_n \to 0$, as $n \to \infty$ which is satisfied by Assumption I. This is particularly important because LASSO-type penalties do not satisfy the oracle property, i.e., they do not ensure that a $\sqrt{n}$-consistent MLE of θ exists which satisfies Part a and Part b. This is because the existence of a $\sqrt{n}$-consistent MLE requires that $\sqrt{n}\lambda_n \longrightarrow \infty$ and the consistency of $\hat{\vartheta}_1$ needs that $\sqrt{n}\lambda_n \longrightarrow 0$. Hence, under a tighter assumption, we show that if such an estimator exists, then it satisfies consistency. Other non-concave penalties like SCAD or HARD, however, can satisfy the oracle property with a proper choice of the tuning parameter.

3.2 Choice of λ_n

Generally the tuning parameters are chosen by cross-validation (Stone, 1974) or generalized cross-validation (Craven and Wahaba, 1979). We should remember that λ_n depends on n. To satisfy the asymptotic properties, we require $\lambda = o\left(\log n/n\right)$. Khalili and Chen (2007) derived a component-wise deviance-based GCV with the above conditions in order to estimate λ. The method, though originally used in regression, also serves well for mixture models. The present paper takes the working sequence $\lambda_n = 1/p$ and studies the behaviour of the LPBIC. The methods proposed by Khalili and Chen (2007), modified for a mixture model, are also considered and provide a range for the values of λ_n. It is observed that for moderately large n ($n \geq 50$), $\lambda_n = 1/p$ falls into that range. For our data analysis (Section 4), we studied the behaviour of LPBIC for different values of λ_n within that range. For illustration, though, a single λ_n is chosen because the behaviour of the LPBIC is uniform over different λ_n values within that range.

4 Data Analysis

4.1 Overview

We analyze two data sets and compare the results using the PBIC to those with the BIC for the PGMM family. The first one is a high-dimensional simulated data set and the second one is a real high-dimensional data set. Although run as cluster analyses, the true group memberships are known in each case and we use the adjusted Rand index (ARI: Rand, 1971; Hubert and Arabie, 1985) to reflect classification agreement. A value of 1 indicates perfect agreement and a value of 0 would be expected under random classification.

4.2 Simulated Data

We generate a simulated p-dimensional Gaussian data set consisting of three groups. We set $\boldsymbol{\mu}_1 = -5.5\mathbf{1}$, $\mathbf{\Sigma}_1$ isotropic; $\boldsymbol{\mu}_2 = 2\mathbf{1}$, $\mathbf{\Sigma}_2$ diagonal; and $\boldsymbol{\mu}_3 = 3\mathbf{1}$, $\mathbf{\Sigma}_3$ full, with $n_1 = 40$, $n_2 = 30$, $n_3 = 30$. We ran simulations for $p \in \{100, 250, 500\}$. LPBIC values are observed for each member of the PGMM family for $G = 1, \ldots, 4$ and $q = 1, 2, 3$. The results (Table 1) show that the PBIC consistently chooses $G = 3$ as p gets larger but that the BIC fails in higher dimensions, choosing a $G = 2$ component model. The associated ARI values (Table 1) confirm that the models selected by the PBIC capture the underlying group structure better than those chosen by the BIC, especially in higher dimensions.

Table 1: Best model chosen by PBIC and BIC for high-dimensional simulated data.

	LPBIC				BIC			
	G	q	Model	ARI	G	q	Model	ARI
$p = 100$	3	3	CUC	0.88	3	3	CUC	0.86
$p = 250$	3	2	CUC	0.82	2	1	CCC	0.62
$p = 500$	3	3	CUC	0.97	2	1	CCC	0.49

The effect of increasing dimension on the performance of the BIC is clear: the BIC chooses fewer mixture components and latent factors, as well as a more parsimonious covariance structure. The LPBIC, however, chooses the same number of components and the same covariance structure each time, and the number of factors does not decrease with p.

Next, we generate 25 simulations of the $p = 500$ dimensional data and study the behaviour of BIC and LPBIC for selecting G and for clustering performance (i.e., ARI). The results (Figure 1) show that LPBIC correctly chooses the number of components ($G = 3$) 23 times but the BIC only selects $G = 3$ four times out of 25. As expected, the BIC tends to choose too few components. The ARIs for models selected using the LPBIC are higher than those selected using the BIC. Out of 25 simulations, the ARI with the LPBIC is higher than that for the BIC in 21 cases, illustrating generally superior clustering performance.

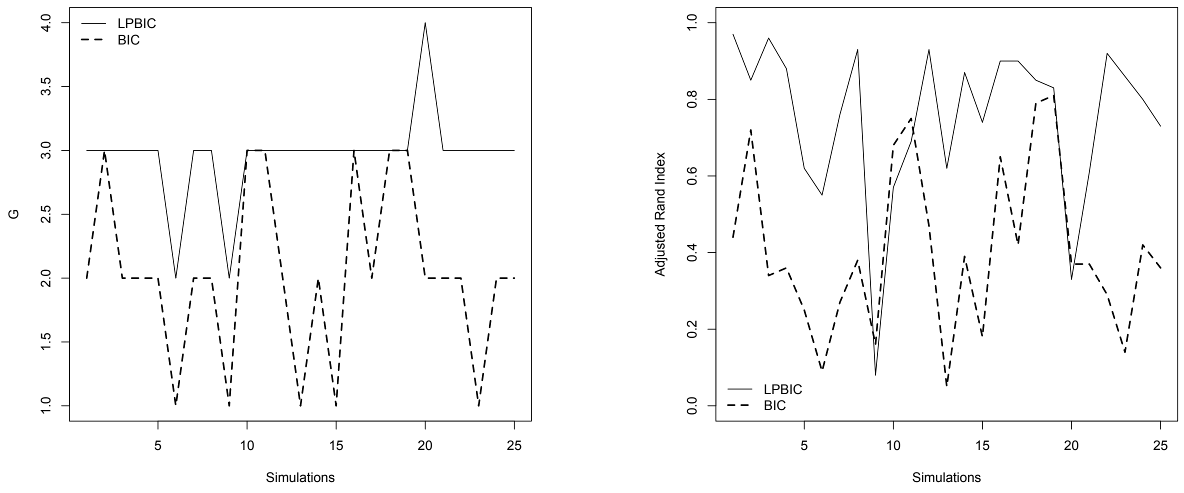

Figure 1: Plot of the performance of LPBIC and BIC for 25 simulations. The left-hand plot shows the selection of number of components by the BIC and LPBIC. The right-hand plot shows the ARIs of the models selected by LPBIC and BIC.

4.3 Leukaemia data

Golub (1999) presented data on two forms of acute leukaemia: acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). Affymetrix arrays were used to collect measurements for 7,129 genes on 72 tissues. There were a total of 47 ALL tissues and 25 with AML. McLachlan et al. (2002) reduced the data set as follows:

1. Genes with expression falling outside the interval $(100, 16000)$ are removed.

2. Genes with expression satisfying max/min ≤ 5 or max-min ≤ 500 are removed.

McNicholas and Murphy (2010) further reduced the number of genes to 2,030 by applying the `select-genes` software (cf. McLachlan et al., 2002). We analyze these 2,030 genes using 20 different random starts for the initial $\hat{z}_{ig}$. We run our approach for $G \in \{1, 2\}$ and $q = 1, \ldots, 6$.

Table 2: Comparison of the performance of LPBIC and BIC for PGMM model selection for the leukaemia data.

	Value	G	q	Model	ARI
BIC	-400394	1	2	CCU	0.29
LPBIC	-391023	2	1	CUC	0.47

Summaries of the models selected by the LPBIC and the BIC, respectively, are given in Table 2. The BIC chooses a CCU model with $G = 1$ component and $q = 2$ factors. The LPBIC chooses a CUC model with $G = 2$ components and $q = 1$ factors. The ARI of the model chosen using LPBIC (0.47) is greater than that for the model chosen using the BIC (0.29). The model selected using the LPBIC misclassifies eleven of the 72 samples (Table 3).

Table 3: Classification table of the best model chosen by LPBIC.

	1	2
ALL	39	3
AML	8	22

5 Discussion

The paper proposes a LPBIC through a penalized likelihood-based approach in the context of parsimonious Gaussian mixture model selection. The approach is mainly intended for the high-dimensional setting, where the BIC has some unattractive problems due to an 'exploding' penalty term for high-dimensional data. Our LPBIC approach does not use the total number of independent parameters to be estimated in its penalty term but, rather, the total number of independent non-zero parameters to be estimated. This has some advantages. Because the likelihood is penalized by a tuning parameter, many of the mean components become 0, thereby reducing the number of independent estimable parameters. The loss of information due to penalizing the likelihood is somehow compensated for by both absolute mean and absolute coefficient of variation of the mean parameters.

The choice of tuning parameters is an important aspect in this scenario because no theoretical result exists which specifies the best choice. Recently, Wang et al. (2007, 2009) proposed some interesting mathematical methods of choosing the tuning parameters without requiring cross-validation. However, their method is most suitable in low-dimensional settings. Herein, we followed an approach close to the one proposed by Fan and Li (2001), though careful modifications have been taken to preserve the asymptotic properties, accounting for the nature of the data.

Our method seems consistent in choosing the right number of clusters for high-dimensional data, as shown through the analysis of real and simulated data. Our analyses suggest that the LPBIC is an improvement over the BIC in the high-dimensional setting. What we lose is the oracle property, because the LASSO may fail to satisfy the consistency, sparsity and asymptotic normality all at the same time. But the LASSO has some computational advantages because of convexity and hence it is preferred over other non-concave penalties.

Of course, the LPBIC is not without its issues. One problem arises by locally approximating the penalty function: if an estimator is shrunken, it stays at 0. Another arises if the initial domain of the estimates does not contain the posterior mode, or even if the posterior mode lies at the boundary of the domain. This second problem, which will lead to failure, is a general problem with the EM algorithm.

Future work will focus on the use of penalties that lead to consistent model selection criteria. We are in particular interested in the adaptive LASSO which leads to the oracle properties. We shall also study the penalization of the variance parameters as it will generate greater parsimony.

Acknowledgements

This work was presented at the Statistical Society of Canada annual meeting in Guelph, Canada (June 2012) and at the MBC^2 meeting in Catania, Italy (September 2012). The authors wish to thank the numerous people who provided helpful comments and feedback at these meetings. This work was supported by a Collaborative Research and Development grant from the Natural Sciences and Engineering Research Council of Canada, a grant-in-aid from Compusense Inc., a Collaborative Research grant from the Ontario Centres of Excellence, and an Early Researcher Award from the Ontario Ministry of Research and Innovation.

References

Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition 28*, 781–793.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Clyde, M., J. Berger, F. Bullard, E. Ford, W. Jeffreys, R. Luo, R. Paulo, and T. Loredo (2007). Current challenges in Bayesian model choice. *Statistical Challenges in Modern Astronomy IV 371*, 224–240.

Craven, P. and G. Wahaba (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematika 31*, 377–403.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fraley, C. and A. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*, 611–631.

Fraley, C. and A. E. Raftery (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification 24*, 155–181.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*, 432–441.

Ghahramani, Z. and G. E. Hinton (1997). The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, University Of Toronto, Toronto.

Golub, T. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*(431), 928–934.

Keirbin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya . The Indian Journal of Statistics. Series A 62*(1), 49–66.

Khalili, A. and J. Chen (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association 102*(479), 1025–1038.

McLachan, G. and K. Basford (1988). *Mixture models: Inference and applications to clustering*. Marcel Dekker Inc.

McLachan, G. and D. Peel (2002). *Finite Mixture Model*. John Wiley & Sons, Inc.

McLachlan, G. J., R. W. Bean, and D. Peel (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics 18*(3), 412–422.

McLachlan, G. J. and D. Peel (2000). Mixtures of factor analyzers. In *Proceedings of the Seventh International Conference on Machine Learning*, San Francisco, pp. 599–606. Morgan Kaufmann.

McNicholas, P. D. and T. B. Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing 18*, 285–296.

McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics 26*(21), 2705–2712.

McNicholas, P. D., T. B. Murphy, A. F. McDaid, and D. Frost (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis 54*(3), 711–723.

Pan, W. and X. Shen (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research 8*, 1145–1164.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*, 846–850.

Shibata, R. (1986). Consistency of model selection and parameter estimation. *Journal of Applied Probability 23*, 127–141.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society: Series B 36*, 111–147.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B 58*, 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B 67*, 91–108.

Tipping, T. E. and C. M. Bishop (1999). Mixtures of probabilistic principal component analysers. *Neural Computation 11*(2), 443–482.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions.* Chichester: John Wiley & Sons.

Wang, H., L. Bo, and C. Ling (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B 971*(3), 671–683.

Wang, H., L. Runze, and C.-L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*(3), 553–568.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

A Derivation of LPBIC

To derive the LPBIC, we closely follow the derivation of the usual BIC. We have to maximize (3). Using (4), the second term becomes

$$n\lambda_n \sum_{g=1}^{G} \int \pi_g \sum_{j=1}^{p} |\mu_{gj}| \mathrm{d}\pi_g = n\frac{\lambda_n}{G} \sum_{g=1}^{G} \sum_{j=1}^{p_g} \left[|\hat{\mu}_{gj}| + \frac{1}{2} \frac{\operatorname{sign}(\hat{\mu}_{gj})}{\hat{\mu}_{gj}} (\mu_{gj}^2 - \hat{\mu}_{gj}^2) \right],$$

where p_g is the number of non-zero mean components in class g. Under the assumption made in Section 2, $\boldsymbol{\mu}_g$ is at most p_g dependent, and the Weak Law of Large Numbers holds. In a large-p setting, $\sum_{g=1}^{G} p_g$ is a large number and so $\sum_{g=1}^{G} \sum_{j=1}^{p_g} \left(\mu_{gj}^2 - \hat{\mu}_{gj}^2\right) / \sum_{g=1}^{G} p_g \xrightarrow{P} \sum_{g=1}^{G} \sum_{j=1}^{p_g} \left(I(\hat{\mu}_g)^{-1}\right)_{jj} / \sum_{g=1}^{G} p_g$. Thus the second term becomes

$$\frac{n\lambda}{G} \sum_{k=1}^{G} \sum_{j=1}^{p_g} \left(|\hat{\mu}_{gj}| + \frac{\left(I(\hat{\mu}_g)^{-1}\right)_{jj}}{|\hat{\mu}_{gj}|} \right).$$

The first term, using Taylor's expansion, is

$$\begin{aligned} &\int \exp\left[\log \mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}) \mathcal{G}(\boldsymbol{\vartheta})\right] \mathrm{d}\boldsymbol{\Theta} \\ &\quad = \int \exp\left[\log \mathcal{L}(\hat{\boldsymbol{\vartheta}} \mid \mathbf{x}) \mathcal{G}(\hat{\boldsymbol{\vartheta}}) + (\boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}}) \frac{\partial \log \mathcal{L}(\boldsymbol{\vartheta}) \mathcal{G}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} - \frac{1}{2} (\boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}})^T \mathcal{H}_{\hat{\boldsymbol{\vartheta}}} (\boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}}) \right] \mathrm{d}\boldsymbol{\vartheta}, \end{aligned}$$

where $\mathcal{H}$ is the second derivative matrix of $\log \mathcal{L}(\boldsymbol{\vartheta}) \mathcal{G}(\boldsymbol{\vartheta})$. Because $\hat{\boldsymbol{\vartheta}}$ is derived maximizing the penalized likelihood, the second term within the integral becomes $(\boldsymbol{\vartheta} - \hat{\boldsymbol{\vartheta}}) \partial \varphi_n(\boldsymbol{\mu}) / \partial \boldsymbol{\vartheta}$, where $\varphi_n(\boldsymbol{\mu})$ is the LASSO penalty function. Using (4), the mean-value theorem and the fact that the $\boldsymbol{\vartheta}$ values are close to $\hat{\boldsymbol{\vartheta}}$, the second term within the integral is $n\lambda_n / G \sum_{g=1}^{G} \sum_{j=1}^{p_g} \operatorname{sign}(\mu_{gj})$.

The third term within the integral similarly becomes $1/2(\tilde{\boldsymbol{\vartheta}} - \hat{\tilde{\boldsymbol{\vartheta}}})' \mathcal{H}_{\hat{\tilde{\boldsymbol{\vartheta}}}} (\tilde{\boldsymbol{\vartheta}} - \hat{\tilde{\boldsymbol{\vartheta}}})$, where $\tilde{\boldsymbol{\vartheta}}$ is the set of non-zero parameters and $\hat{\tilde{\boldsymbol{\vartheta}}}$ is their estimate. Using Laplace approximation on $\mathcal{H}$ and applying the Weak Law of Large Numbers, as in the usual BIC, we arrive at $\log \mathcal{L}(\hat{\boldsymbol{\vartheta}} \mid \mathbf{x}) - 1/2 \tilde{\rho} \log n$, where $\tilde{\rho} = \dim(\hat{\tilde{\boldsymbol{\vartheta}}})$. This, combined with the second term of (3), gives (5).

B Proof of the Asymptotic Property of LPBIC

First, suppose the true number of clusters G is known with the corresponding parameter $\boldsymbol{\vartheta}$. Let the true parameter be $\boldsymbol{\vartheta}_0$. Let $\hat{\boldsymbol{\vartheta}}$ be an arbitrary estimate of $\boldsymbol{\vartheta}$. Let $\tilde{\rho}_0$ and $\tilde{\rho}_1$ be the corresponding number of non-zero parameters and $\lambda_n^{(0)}$ and $\lambda_n^{(1)}$ be the corresponding tuning parameters. We first prove that, for an arbitrary estimate $\hat{\boldsymbol{\vartheta}}$ satisfying $\| \hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0 \| = \mathcal{O}(n^{\kappa})$, $\text{LPBIC}(\hat{\boldsymbol{\vartheta}}_1, \hat{\boldsymbol{\vartheta}}_2) - \text{LPBIC}(\hat{\boldsymbol{\vartheta}}_1, \mathbf{0}) \leq 0$ as $n \longrightarrow \infty$. We note that

$$\text{LPBIC}(\hat{\boldsymbol{\vartheta}}_1, \hat{\boldsymbol{\vartheta}}_2) - \text{LPBIC}(\hat{\boldsymbol{\vartheta}}_1, \mathbf{0}) = 2l(\hat{\boldsymbol{\vartheta}}_1, \hat{\boldsymbol{\vartheta}}_2 \mid \mathbf{x}) - 2l(\hat{\boldsymbol{\vartheta}}_1, \mathbf{0} \mid \mathbf{x}) - \left[\Lambda(\hat{\boldsymbol{\vartheta}}_1, \hat{\boldsymbol{\vartheta}}_2) - \Lambda(\hat{\boldsymbol{\vartheta}}_1, \mathbf{0})\right],$$

where $l = \log \mathcal{L}$ and Λ is the penalty part of LPBIC. Using the mean-value theorem,

$$l(\hat{\boldsymbol{\vartheta}}_1, \hat{\boldsymbol{\vartheta}}_2 \mid \mathbf{x}) - l(\hat{\boldsymbol{\vartheta}}_1, \mathbf{0} \mid \mathbf{x}) = \left[\frac{\partial l(\hat{\boldsymbol{\vartheta}}_1, \xi)}{\partial \boldsymbol{\vartheta}_2}\right]' \hat{\boldsymbol{\vartheta}}_2,$$

where $\| \xi \| \leq \| \hat{\boldsymbol{\vartheta}}_2 \| = \mathcal{O}(n^{\kappa})$. Also,

$$\begin{aligned} \left\| \frac{\partial l(\hat{\boldsymbol{\vartheta}}_1, \xi)}{\partial \boldsymbol{\vartheta}_2} - \frac{\partial l(\boldsymbol{\vartheta}_0, \mathbf{0})}{\partial \boldsymbol{\vartheta}_2} \right\| &\leq \left\| \frac{\partial l(\hat{\boldsymbol{\vartheta}}_1, \xi)}{\partial \boldsymbol{\vartheta}_2} - \frac{\partial l(\hat{\boldsymbol{\vartheta}}_1, \mathbf{0})}{\partial \boldsymbol{\vartheta}_2} \right\| + \left\| \frac{\partial l(\hat{\boldsymbol{\vartheta}}_1, \mathbf{0})}{\partial \boldsymbol{\vartheta}_2} - \frac{\partial l(\boldsymbol{\vartheta}_0, \mathbf{0})}{\partial \boldsymbol{\vartheta}_2} \right\| \\ &\leq \sum_{i=1}^{n} M_2(z_i) \left[||\xi|| + \left\| \hat{\boldsymbol{\vartheta}}_1 - \boldsymbol{\vartheta}_0 \right\| \right] = \left[||\xi|| + \left\| \hat{\boldsymbol{\vartheta}}_1 - \boldsymbol{\vartheta}_0 \right\| \right] \mathcal{O}(n) = \mathcal{O}\left(n^{\kappa+1}\right) \end{aligned} \tag{6}$$

from Assumption II. Also, from the last part of the first line of (6), which is of order $\mathcal{O}\left(n^{\kappa+1}\right)$, we can conclude that $\partial l(\boldsymbol{\vartheta}_0, \mathbf{0}) / \partial \boldsymbol{\vartheta}_2$ is of order $\mathcal{O}\left(n^{\kappa+1}\right)$, as is $\partial l(\hat{\boldsymbol{\vartheta}}_1, \xi)/\partial \boldsymbol{\vartheta}_2$. Therefore, from these order assessments, we conclude that

$$l\left(\tilde{\boldsymbol{\vartheta}}_1, \tilde{\boldsymbol{\vartheta}}_2\right) - l\left(\tilde{\boldsymbol{\vartheta}}_1, \mathbf{0}\right) = \mathcal{O}\left(n^{\kappa+1}\right) \sum_{g=1}^{G} \sum_{j=p_g+1}^{p} \hat{\mu}_{gj},$$

where p_g is defined as in (4).

For the part $\Lambda(\hat{\boldsymbol{\vartheta}}_1, \boldsymbol{\vartheta}_2) - \Lambda(\hat{\boldsymbol{\vartheta}}_1, \mathbf{0})$, note that

$$\frac{2n\lambda_n}{G} \sum_{g=1}^{G} \sum_{j=1}^{p_k} \left[|\hat{\mu}_{gj}| + \frac{\left(I(\hat{\boldsymbol{\mu}}_g)^{-1}\right)_{jj}}{|\hat{\mu}_{gj}|} - \text{sign}\left(\tilde{\mu}_{gj}\right) \right] = \mathcal{O}\left(n^{\alpha+1}\right) \lambda_n$$

because the summation part is some constant times $p = \mathcal{O}(n^{\alpha})$, using Assumption I. We also have $(\tilde{\rho}_1 - \tilde{\rho}_0) \log n = \sum_{g=1}^{G} (p - p_g) \log n = \mathcal{O}(n^{\alpha}) \log n$. Hence,

$$\text{LPBIC}(\hat{\boldsymbol{\vartheta}}_1, \hat{\boldsymbol{\vartheta}}_2) - \text{LPBIC}(\hat{\boldsymbol{\vartheta}}_1, \mathbf{0}) = \mathcal{O}(n^{\kappa+1}) \sum_{g=1}^{G} \sum_{j=p_g+1}^{p} \hat{\mu}_{gj} - \mathcal{O}(n^{\alpha}) \log n - (\lambda_n^{(1)} - \lambda_n^{(0)}) \mathcal{O}(n^{\alpha+1}).$$

The first term of the above expression is $\mathcal{O}\left(n^{\kappa+1}\right) \sum_{g=1}^{G} \sum_{j=p_g+1}^{p} \hat{\mu}_{gj} = \mathcal{O}\left(n^{2\kappa+1}\right)$. Using Assumption I, i.e., that $\lambda_n = o(\log n/n)$, and by order comparison, we can conclude that the leading terms in the above expression are $\mathcal{O}\left(n^{2\kappa+1}\right)$ and $\mathcal{O}(n^{\alpha}) \log n$. Because $\alpha > 2\kappa + 1$, $\text{LPBIC}\left(\hat{\boldsymbol{\vartheta}}_1, \hat{\boldsymbol{\vartheta}}_2\right) - \text{LPBIC}\left(\hat{\boldsymbol{\vartheta}}_1, \mathbf{0}\right) \leq 0$ as $n \longrightarrow \infty$.

Now, let $\tilde{\boldsymbol{\vartheta}} = \left(\tilde{\boldsymbol{\vartheta}}_1, \tilde{\boldsymbol{\vartheta}}_2\right)$ be an estimate of $\boldsymbol{\vartheta}$ such that $(\tilde{\boldsymbol{\vartheta}}_1, \mathbf{0})$ is a maximizer of $\text{LPBIC}(\boldsymbol{\vartheta}_1, \mathbf{0})$ satisfying $\| \tilde{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0 \| = \mathcal{O}(n^{\kappa})$. It suffices to show that in the neighbourhood $\| \boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0 \| = \mathcal{O}(n^{\kappa})$, $\text{LPBIC}(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2) - \text{LPBIC}\left(\tilde{\boldsymbol{\vartheta}}_1, \mathbf{0}\right) < 0$ with probability tending to 1 as $n \to \infty$. We note that

$$\text{LPBIC}(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2) - \text{LPBIC}(\tilde{\boldsymbol{\vartheta}}_1, \mathbf{0}) = [\text{LPBIC}(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2) - \text{LPBIC}(\boldsymbol{\vartheta}_1, \mathbf{0})] + [\text{LPBIC}(\boldsymbol{\vartheta}_1, \mathbf{0}) - \text{LPBIC}(\tilde{\boldsymbol{\vartheta}}_1, \mathbf{0})],$$

where $\text{LPBIC}(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2) - \text{LPBIC}(\boldsymbol{\vartheta}_1, \mathbf{0}) \leq 0$ with probability tending to 1 (by the previous result) and $\text{LPBIC}(\boldsymbol{\vartheta}_1, \mathbf{0}) - \text{LPBIC}(\tilde{\boldsymbol{\vartheta}}_1, \mathbf{0}) \leq 0$ with probability tending to 1 since $(\tilde{\boldsymbol{\vartheta}}_1, \mathbf{0})$ is a maximizer of $\text{LPBIC}(\boldsymbol{\vartheta}_1, \mathbf{0})$. Thus $(\tilde{\boldsymbol{\vartheta}}_1, \mathbf{0})$ maximizes $\text{LPBIC}(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2)$ with probability tending to 1 as $n \to \infty$. Hence we conclude that $\text{P}\left(\tilde{\boldsymbol{\vartheta}}_2 = \mathbf{0}\right) \longrightarrow 1$ as $n \to \infty$. Hence the proof.

The case of unknown clusters can be similarly proved. If the estimated number of components is G_1 and the true number is G, then the estimated parameter corresponding to G_1 is, say, $\hat{\boldsymbol{\vartheta}}$. We can again decompose $\hat{\boldsymbol{\vartheta}}$ as $(\hat{\boldsymbol{\vartheta}}_1, \hat{\boldsymbol{\vartheta}}_2)$ and similarly show that $\hat{\boldsymbol{\vartheta}}_2 \longrightarrow 0$ in probability. Here $\hat{\boldsymbol{\vartheta}}_1$ comprises of the clusters belonging to $\boldsymbol{\vartheta}_0$.